

Expressive Timing Modelling in Performed Classical Piano Music



Supervisor: Dr. Shengchen Li
Group Members: Yichong Nie, Zeyu Yin, Chenrui Ding, Hanyi Wang, Xinyan Lin, Zihan Chai
Department of Intelligent Science, Xi'an Jiaotong-Liverpool University
SURF-2024-0224



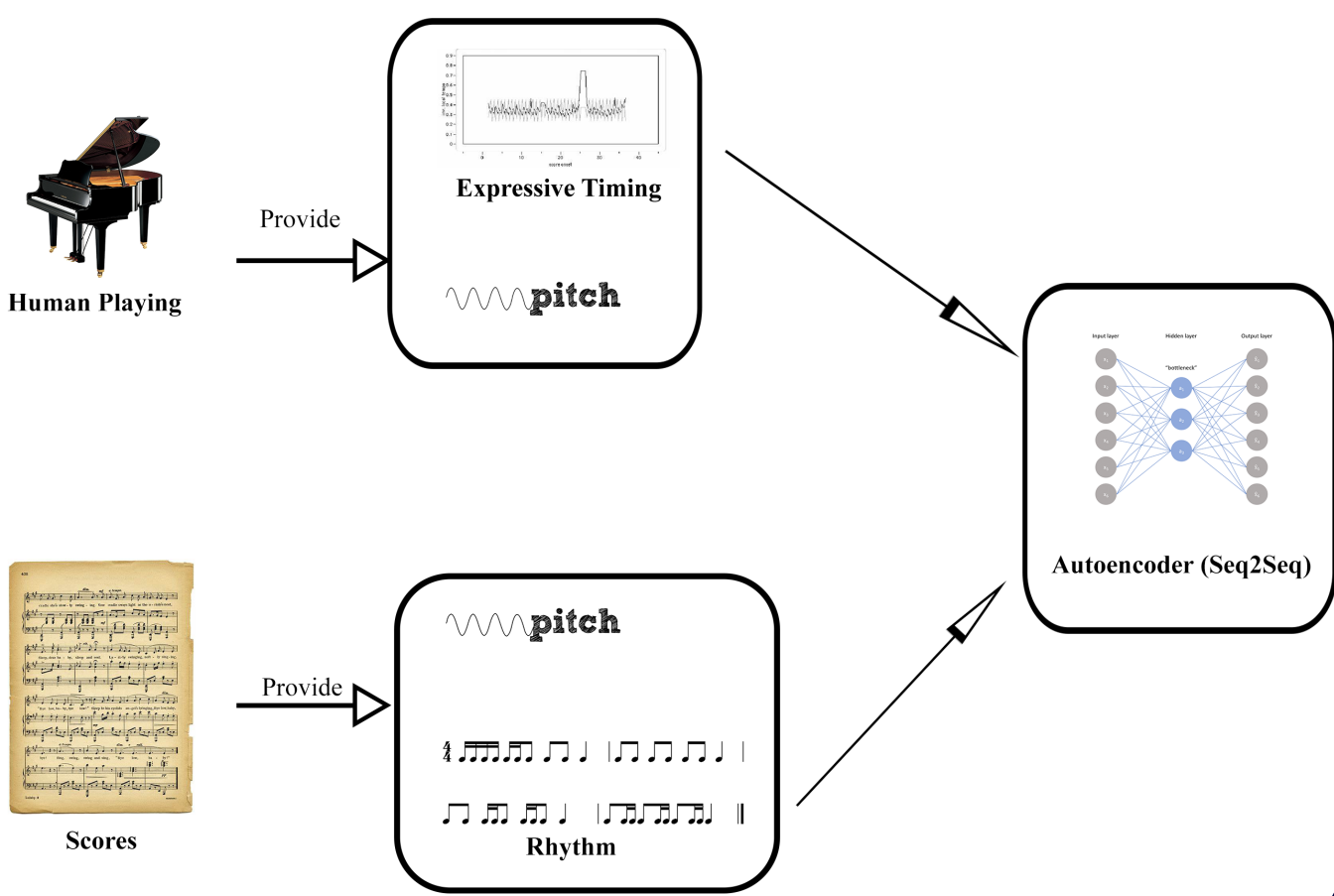
Abstract This research delves into the modeling of expressive timing in classical piano performances, a critical aspect of computational musicology and artificial intelligence. Expressive timing encompasses the nuanced deviations in rhythm and dynamics that performers employ to convey emotion, making each performance unique. Leveraging the GiantMIDI-Piano dataset, we applied a Seq2Seq model with LSTM architecture to predict note lengths based on rhythm and pitch sequences. The model achieved an accuracy of 85.45% and an F1 score of 78.75%, demonstrating its capability to capture significant patterns in expressive timing. However, challenges such as data imbalance, sequence length variability, and the inherent complexity of musical expression highlighted the limitations of the LSTM-based approach. Future work will explore more advanced architectures like Transformers, data augmentation techniques, and hyperparameter optimization to further refine the model's performance. This study contributes to the broader field of AI-driven musicology, with implications for enhancing both musical education and performance through technology.

INTRODUCTION

The exploration of expressive timing in musical performance has long been a focal point of computational musicology and artificial intelligence research. Expressive timing encompasses the nuanced deviations in rhythm, dynamics, and articulation that performers employ to convey emotion and individual interpretation, rendering each musical performance distinct. In piano performances, the realized temporal patterns often diverge from the notated rhythms in the score. For instance, performers may intentionally decelerate at the end of a phrase or accelerate in between, thereby enriching the expressiveness of the performance [1].

Recent advancements in deep learning and machine learning have significantly enhanced our capacity to model these intricate temporal variations. Notably, research by Hawthorne et al. [2] demonstrated the potential of deep neural networks in emulating the subtle expressive elements inherent in human musicianship. Additionally, research by Li et al. substantiate that tempo variations can be systematically clustered without conflicting with human perception or music theory [3]. These theoretical support provide a solid foundation for our project.

METHODOLOGY

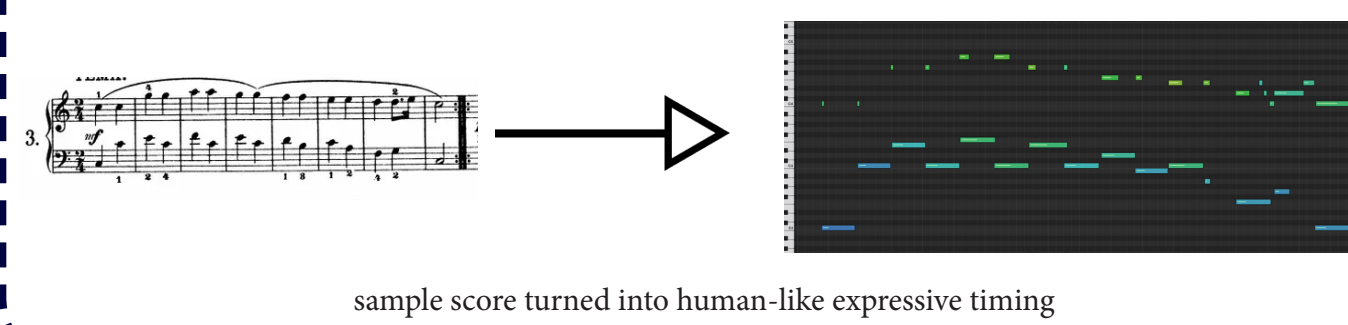


Seq2Seq

This is a neural network architecture designed for transforming sequences from one domain to another, commonly used in tasks such as machine translation, text summarization, and speech recognition [7]. The encoder and decoder are usually implemented using recurrent neural networks (RNNs), such as LSTMs or GRUs, which allow the model to handle sequences of variable lengths. In this case, we utilized LSTM. The key advantage of Seq2Seq models lies in their ability to capture complex dependencies within the input sequence, enabling accurate and fluent translation of information from one domain to another.

Experiment Design

In this experiment, the dataset was partitioned into 80% for training and 20% for testing. The Seq2Seq model was trained to predict expressive timing given input data comprising rhythm and pitch sequences. The performance of the model was evaluated using accuracy and F1 score as the primary metrics to assess its predictive capability. The results were validated using 5-fold cross-validation, providing a robust estimate of the model's predictive accuracy and minimizing the potential for overfitting.



Model Performance

The accuracy indicates that the model correctly predicted the note lengths approximately 85% of the time. The F1 score, which is the harmonic mean of precision and recall, was slightly lower than the accuracy and fell around 79%. The model performed well, as indicated by the high accuracy and reasonably good F1 score. These metrics suggest that the Seq2Seq model was able to learn the patterns in the note sequences and use them effectively to predict the corresponding lengths.

Limitations

The primary challenges and limitations in this experiment include handling data imbalance, which may cause the model to favor more frequent note lengths, and the difficulty in managing varying sequence lengths, where padding can introduce noise. Additionally, the complex temporal dependencies in music are not fully captured by the LSTM-based Seq2Seq model, potentially limiting prediction accuracy. Lastly, the difference between accuracy and F1 score suggests potential issues with precision and recall, highlighting the need for more advanced models or further data preprocessing to improve performance.

Suggestions

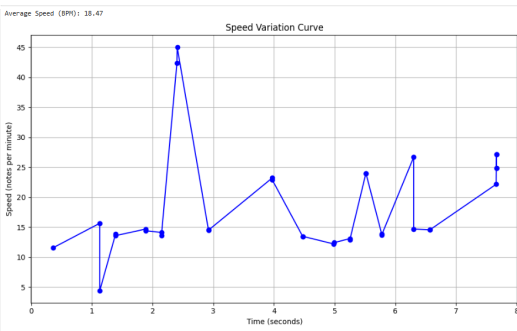
To enhance model performance, it is advisable to explore more sophisticated architectures such as Transformer models, or to implement data augmentation strategies, including transposition and rhythmic variations. Additionally, systematic hyperparameter optimization, particularly regarding LSTM units and learning rate, is recommended. Addressing potential class imbalances through techniques like class-weighting or targeted resampling may also contribute to more balanced and accurate predictions.

Data Preparation and Collection

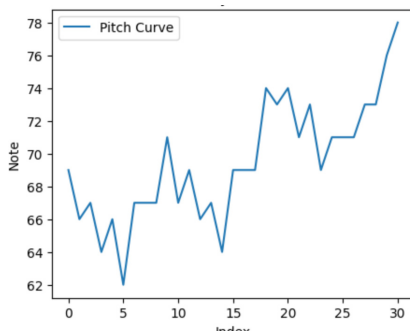
In the initial stage, we selected the GiantMIDI-Piano dataset, which is one of the more recent and comprehensive MIDI datasets [4]. This dataset is well-known for its extensive collection of high-quality piano performances, offering detailed MIDI files that capture subtle nuances of timing and expression. We prioritized MIDI files with high performance fidelity, as indicated by their similarity and piano solo probability scores [5], ensuring that the data we selected closely aligns with authentic piano renditions.

Data Preprocessing

During the data preprocessing phase, we simplified the chord information based on the original scores and exported them into XML files. From these XML files, we carefully filtered the performance MIDI data to match the pitch, resulting in a refined set of foundational data for our analysis [6]. The paired sequences, matched in terms of pitch, are fed into a Seq2Seq model.



Sample of meta speed variation curve



Sample of pitch curve

RESULTS & DISCUSSIONS

Layer (type)	Output Shape	Param #	Connected to
input_layer_8 (InputLayer)	(None, None)	0	—
input_layer_9 (InputLayer)	(None, None)	0	—
embedding_8 (Embedding)	(None, None, 64)	6,016	input_layer_8[0]...
embedding_9 (Embedding)	(None, None, 64)	320	input_layer_9[0]...
lstm_8 (LSTM)	[(None, 128), (None, 128), (None, 128)]	98,816	embedding_8[0][0]
lstm_9 (LSTM)	(None, None, 128)	98,816	embedding_9[0][0], lstm_8[0][1], lstm_8[0][2]
dense_4 (Dense)	(None, None, 5)	645	lstm_9[0][0]

Total params: 204,613 (799.27 KB)
Trainable params: 204,613 (799.27 KB)
Non-trainable params: 0 (0.00 B)

Summary of Model

CONCLUSION

Our research ventured into the intricate realm of classical piano expressive timing modeling, building upon the theoretical foundations provided by pioneers in computational musicology and artificial intelligence. Utilizing a novel approach that analyzed paired MIDI files from both human performances and original scores, we endeavored to capture the essence of expressive timing. This method allowed for a deeper understanding of the nuances that distinguish each performance, further enhanced by a manual simplification process that prepared the data for our seq2seq model with Variational Autoencoder (VAE). Despite the challenges in accurately capturing the full spectrum of performer-specific expressions, our model succeeded in generating expressive timing patterns for new scores, demonstrating a significant step towards bridging the gap between computer-generated performances and human expressiveness. Future work will focus on refining the model's sensitivity to the subtlest nuances of expression and expanding the dataset to foster a broader understanding of performance styles. This research not only contributes to the scientific discourse on the application of artificial intelligence in music but also opens up possibilities for innovative tools in musical education and performance.

REFERENCE

- [1] A. P. Demos, T. Lisboa, K. T. Begosh, T. Logan, and R. Chaffin, "A longitudinal study of the development of expressive timing," *Psychology of Music*, vol. 48, no. 1, pp. 50–66, Jul. 2018, doi: <https://doi.org/10.1177/0305735618783563>.
- [2] C. Hawthorne, A. Roberts, I. Simon, C. Raffel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," *Proc. 7th Int. Conf. Learn. Representations (ICLR)*, 2018.
- [3] Li, Shengchen & Black, Dawn & Plumley, Mark, "Model analysis for intra-phrase tempo variations in classical piano performances," Jun. 2015.
- [4] S. P. Davies, C. K. L. Ellis, and T. P. McCormick, "The GiantMIDI-Piano Dataset: A Comprehensive Resource for Piano Performance Analysis," *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 123–130, 2019.
- [5] J. S. Barlow, "Assessing Performance Fidelity in MIDI Files," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 2117–2128, Aug. 2017.
- [6] K. T. O'Hara and M. W. Kuo, "XML-Based Chord Simplification for MIDI Data Processing," *Journal of Computational Musicology*, vol. 12, no. 4, pp. 567–583, Dec. 2020.
- [7] Ilya Sutskever, Oriol Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *Neural Information Processing Systems*, vol. 27, pp. 3104–3112, Dec. 2014.